Attorney Docket No.: 19488-000710US

PATENT APPLICATION

HIGH-THROUGHPUT TRANSCRIPTOME ANALYSIS

Inventors: Thorsten Melcher, a citizen of Germany
residing at 1241 8th Avenue No. 2
San Francisco, CA

K.C. McFarland, a citizen of the United States
residing at 5219 Cowell Blvd.
Davis, CA 95616

Assignee: AGY Therapeutics, Inc.
290 Utah Ave.
South San Francisco, CA 94080

Entity: Small

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 650-326-2400

HIGH-THROUGHPUT TRANSCRIPTOME ANALYSIS

BACKGROUND

It is estimated that while over 100,000 genes are expressed by a mammalian genome, only a fraction are expressed in any particular cell or tissue. Gene expression patterns, especially as reflected in the abundance of mRNAs, vary according to cell or tissue type, with developmental or metabolic state, in response to insult or injury, and as a consequence of other genetic and environmental factors. Moreover, the pattern of expression changes in a dynamic fashion over time with changes in cell state and environment. The term "transcriptome" has been coined to describe the set of all genes expressed, at any given time, under defined conditions in a given tissue (Velculescu et al., 1997, *Cell* 88:243-51).

The detection of changes to the transcriptome can provide useful information regarding the identity of genes and gene products important in development, drug response, and, particularly, human disease processes. However, methods now used for identifying changes in the transcriptome suffer from a variety of deficiencies, e.g., they are expensive, require relatively large quantities of starting material, and/or do not efficiently identify low abundance transcripts important in mediating cell processes. Thus, new and efficient methods for high-throughput analysis of gene expression are needed.

BRIEF SUMMARY OF THE INVENTION

In one aspect, the invention provides a method for selecting clones for analysis by (a) preparing double-stranded cDNA (dscDNA) corresponding to mRNA from each of a pair of related tissues or cells (with one member of the pair designated the driver-tissue and the second member of the pair designated the tester-tissue); (b) using the dscDNA to prepare a driver-normalized cDNA library, a tester-normalized cDNA library, a driver-subtracted cDNA library, and a tester-subtracted cDNA library; (c) hybridizing clones from each of the libraries with detectably labeled cDNA probes corresponding to mRNA from one or both of the related tissues or cells; (d) selecting low intensity signal clones from the driver-normalized cDNA library hybridized with cDNA probe from the driver tissue; (e) selecting low intensity signal clones from the tester-normalized cDNA library hybridized with cDNA probe from the tester tissue; (f) selecting high-signal ratio

clones from the driver-subtracted cDNA library hybridized with cDNA probe corresponding to mRNA from both of the related tissues; and, (g) selecting high signal ratio clones from the tester-subtracted cDNA library hybridized with cDNA probe corresponding to mRNA from both of the related tissues.

In various embodiments of the invention, the tester and driver tissues are from rat, mouse, human or nonhuman primate. In some embodiments, the pair of tissues are related as diseased tissue and healthy tissue, for example, diseased and healthy tissue from brain. In some embodiments, the healthy and diseased tissues are from an animal or tissue culture model of stroke, Alzheimer's disease, neuropathy, or the like.

In one aspect, the invention provides a method of identifying redundant clones in a cDNA library by (a) "prior sampling" or identifying at least one redundant clone in a first portion of the cDNA library; (b) obtaining an isolated polynucleotide corresponding to the redundant clone; (c) hybridizing a detectably-labeled probe to an array of clones from the cDNA library, where the hybridizing is done in the presence and absence of the isolated polynucleotide obtained in (b); (d) comparing the hybridization signal obtained for each arrayed clone in the presence and absence of the isolated polynucleotide; and, (e) identifying clones for which the hybridization signal produced is different in the presence and absence of the isolated polynucleotide as redundant clones. The redundant clone(s) in the first portion of the cDNA library may be identified by determining the sequence or partial sequence of a subset of clones in the library, and comparing these sequences with each other or with a database of sequences to identify multiply represented sequences. In one embodiment, the redundant clone(s) is identified by comparing the sequences of at least 100 clones from a portion of the cDNA library. In various embodiments the isolated polynucleotide(s) is unlabeled, or is detectably labeled, e.g., with a label that can be distinguished from the detectably labeled probe.

In a related embodiment, the invention provides a method of identifying previously characterized clones in a cDNA library by (a) obtaining an isolated polynucleotide corresponding to previously identified clones; (b) hybridizing a detectably labeled probe to an array of clones from the cDNA library in the presence and absence of the isolated polynucleotide obtained in (a); (d) comparing the hybridization signal obtained for each arrayed clone in the presence and absence of the isolated polynucleotide; and, (e) identifying clones for which the hybridization signal produced is different in the presence and absence of the isolated polynucleotide as previously characterized clones.

In one aspect, the invention provides an improved method of making a normalized or subtracted cDNA library by (a) obtaining double-stranded cDNA (dscDNA) corresponding to mRNA from a tissue or cell; (b) restricting a portion of the dscDNA with one restriction enzyme and restricting another portion of the dscDNA with a different restriction enzyme. The two restriction enzymes are selected to produce restriction fragments having a predicted average fragment size of between about 100 and about 500 basepairs and to generate fragments of about the same length (i.e., usually within about 150 basepairs in length of each other). The predicted average length of fragments generated by restriction digests of characteristic sequences (e.g., mRNAs from rat, mouse, human, or non-human primates) may be determined by inspection or computer analysis of sequences found in public databases such as Genbank.

In one aspect, the invention provides methods for identifying improved methods for preparing normalized and/or subtracted cDNA libraries. The method involves making libraries from the same tester and driver RNA but using different methods (e.g., varying parameters such as ratio of tester to driver cDNA mixed during subtraction) and comparing the quality of a two different subtracted cDNA libraries, by (a) obtaining a first subtracted cDNA library and a second subtracted cDNA library, wherein each library is prepared from the same tester and driver RNAs; (b) preparing detectably labeled probe from DNA from each library; (c) hybridizing said probe from each library to an array of immobilized polynucleotides, wherein at least a plurality of said polynucleotides have the sequence of genes that are differentially expressed in the tester RNA compared to the driver RNA, and detecting the hybridization of the probe to the immobilized polynucleotides; (d) identifying at least one immobilized polynucleotide having a sequence that is differentially expressed in the tester RNA compared to the driver RNA and comparing the level of hybridization of probe from the first subtracted cDNA library to said polynucleotide with the level of hybridization of probe from the second subtracted cDNA library to said polynucleotide, wherein, the library having the higher level of hybridization of probe to said polynucleotide is identified as a higher quality library; (e) determining that all clones in the subtracted-normalized library provide approximately equivalent hybridization signals.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 shows duplicate arrays probed using the "knock-down" methods of the invention. Arrows show (A) presence of hybridization signal (triplicate spots) and (B) reduction of signal due to inclusion of knock-down polynucleotide during hybridization. This figure shows a portion (detail) of a larger array.

Figure 2. Clones representing a group that are upregulated in Rsa I, 6h (tester) as opposed to Rsa I, 0h (driver) and are of low hybridization signal (=low abundance) in tester and driver are increased in their signal (abundance) under condition of Library ID "F" (normalized tester-subtracted) and PCR cycles =21, 23, 25, 27. Libraries (L) and numbers of amplification steps in the second PCR cycle (N) are indicated by the shorthand "LN." For example, "A21" encodes a description of Library ID "A" with second PCR cycle process length of 21 cycles.

Figure 3. Clones representing a group that are upregulated in Rsa I, 6h (tester) as opposed to Rsa I, 0h (driver) and are of low hybridization signal (=low abundance) in tester and driver are increased in their signal (abundance) under condition of Library IDs "C" through "F" (normalized tester-subtracted), "H" through "K" (normalized driver-subtracted) and PCR cycles =25. Clones from Library IDs "A" and "B" are essentially unchanged.

Figure 4. Clones representing groups that are upregulated in Rsa I, 6h (tester) as opposed to Rsa I, 0h (driver) and are of low, medium or high tester hybridization signal are normalized in their signal under condition of Library ID "B".

DETAILED DESCRIPTION

The present invention provides methods for efficiently identifying and characterizing genes that play important roles in cellular processes such as aging and development, response to environmental challenges (e.g., injury or drug exposure), and the like. In one aspect, the invention provides methods for identifying novel genes important in cell processes such as those involved in the development of human disease. In another aspect, the invention provides methods linking expression of known genes to specific changes in cell state (e.g., disease, injury, and response to disease or injury), i.e., functional analysis of novel and known genes.

Specifically, the methods disclosed herein permit the rapid and economical generation of "libraries" of differentially expressed and low abundance sequences likely to play roles in pathogenesis and treatment of human disease. Importantly, the methods of the invention are well suited to use with very small amounts of tissue. This advantage permits comprehensive libraries to be produced even when small amount of starting material is available.

Certain aspects of the invention will now be described in detail.

I. Definitions and General Cloning Methods

A. The term "tissue," as used herein in the context of a source of mRNA and cDNA, refers to any aggregation of morphologically or functionally related cells, or cell systems, and thus includes cells (including *in vitro* cultured cells), tissues, organs, and the like.

B. The term "library" as used herein, refers to a collection of polynucleotides (usually in the form of double-stranded cDNA) derived from mRNA of a particular tissue. The polynucleotides of a library may be, but are not necessarily, cloned into a vector.

C. General cloning methods: Methods for manipulation of nucleic acids, including methods useful for carrying out the present invention, are well known and described in references such as Ausubel et al., 1999, CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, Greene Publishing and Wiley-Interscience, New York Supplement including SUPPLEMENT 46 (April 1999) (hereinafter "Ausubel") and in Sambrook et. al., MOLECULAR CLONING - A LABORATORY MANUAL (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989 (hereinafter referred to as "Sambrook").

II. Preparation Of Libraries

A. General

In one aspect of the invention, cDNA libraries are prepared that are highly enriched for gene sequences likely to play a role in the molecular and cellular pathomechanisms of disease, or which are involved in other important cellular processes. In one embodiment of the invention, four related, or "cognate," libraries are prepared and

selected sequences analyzed. Although, in some embodiments of the invention, fewer than four libraries are prepared, by screening multiple (e.g., four) libraries the coverage of the transcriptome is maximized and the likelihood of identifying low-abundance and differentially-expressed genes is increased. Moreover, by preparing four libraries validation techniques, as described *infra* are facilitated.

B. Tissue Sources

The libraries of the invention are prepared using mRNA from pairs of tissues that are of the same type, but which differ in one major characteristic, such disease state (e.g., diseased & normal brain tissue), age (e.g., adult and fetal liver tissue), exposure to drugs, or other state (e.g., stimulated & unstimulated; activated & unactivated), etc. The tissue source may be human or non-human. Typically the tissues are from a mammal such as a human, non-human primate, rat, or mouse. In some embodiments, the tissues are from an animal or tissue culture model of a human disease, e.g., stroke, Alzheimer's disease, and neuropathy.

Examples of tissue pairs useful for library preparation are shown in Table 1.

TABLE 1

Gene-expression state 1	Gene-expression state 2
Diseased tissue a) hypoxic/ischemic brain b) cirrhotic liver c) tumor d) Alzheimer's brain	Normal tissue a) healthy brain b) healthy liver c) normal tissue d) healthy brain
Drug-exposed tissue a) kainate-injected brain b) Zyprexa®-injected brain c) toxin-stimulated cell line	Non-drug exposed tissue a) saline injected brain b) saline injected brain c) saline stimulated cell line
Age/Tissue Type/etc. a) mature brain b) hippocampus c) neurons	Age/Tissue Type/etc. a) fetal brain b) cortex c) glial cells

Although each of any group of four cognate libraries is prepared using the same tissue pair, the libraries have different properties as a result of differences in their construction. For each set of libraries, one tissue in the pair is designated the "driver tissue" (from which "driver" cDNA may be made) and the second tissue in the pair is designated the "tester" tissue (from which "tester" cDNA may be made). For example, in a pair in the same horizontal row of Table I) the tissue in the first column may be considered the tester and the tissue in the second column may be considered the driver. For purposes of the invention, it is entirely arbitrary which tissue is "driver" and which is "tester."

For ease of reference, the four cognate libraries are referred herein as: (1) driver-normalized, (2) tester-normalized, (3) driver-subtracted, and (4) tester-subtracted. Libraries (1) and (2) are normalized, and thus enriched in sequences corresponding to low abundance transcripts. In a cognate group, Library 1 is made using one tissue of a pair (driver tissue) and Library 2 is made using the specified tester tissue. Libraries (3) and (4) are subtracted (or normalized and subtracted) libraries and thus enriched in sequences that are differentially expressed between pairs of tissue states. Libraries (3) and (4) of a cognate group are made using both tissues in the tissue pair.

Several methods are known form making normalized and/or subtracted cDNA libraries. Although certain methods are described or referred to in Sections II(B)-(E), *infra*, the invention is not limited to embodiments in which these methods are used. For example, the analytical methods described in Section III may be used in combination with a variety of normalization/subtraction approaches.

B. Preparation of Double-Stranded cDNA From Paired Tissue Samples

Double-stranded cDNA (dscDNA) is prepared from tissues using standard protocols, i.e., by reverse transcription of messenger (poly A^+) RNA from a specified RNA source using a primer to produce single stranded cDNA. Methods for isolation of total or poly(A) RNA and for making cDNA libraries are well known in the art, and are described in detail in Ausubel and Sambrook. In one embodiment, the library is made using oligo(dT) primers for first strand synthesis. The single-stranded cDNA is converted into double stranded cDNA (dscDNA) using routine methods (see, e.g., Ausubel *supra*).

C. Restriction Enzyme Digestion

In some embodiments of the invention, the dscDNA from each tissue source is digested with one restriction enzyme or, in an alternative embodiment, the dscDNA from each tissue source is separately digested with two or more restriction enzymes, with different specificities, that cut at a recognition sequences found frequently in the dscDNA. Often, two enzymes are used (and the discussion and examples below will refer to use of two enzymes). As noted, the digestion with each of the two or more enzymes is carried out separately (e.g., in separate reaction tubes). The digested fragments may be combined later for further processing.

The dual digestion steps allow the efficient generation of libraries that are more comprehensive (e.g., containing more different species of expressed or differentially expressed species) than libraries made by other methods. The digestion is intended, in part, to generate fragments in a size range that allows efficient hybridization during the annealing steps of library construction. Only fragments of the target size range will efficiently anneal under the conditions used, and non-annealing molecules are excluded from amplification or cloning in some embodiments of the invention. A further advantage of the dual digestion steps is that by digesting with multiple (e.g., 2) enzymes with different specificities as taught herein, the resulting libraries are more comprehensive.

According to the invention, the restriction enzymes used are selected that will produce a calculated (or "predicted") average fragment size of between about 100 and about 500 basepairs, preferably about 300-500 basepairs (e.g., an average length of between 300 bases and 500 bases). In addition, the two or more different enzymes should produce fragments of similar lengths (e.g., so that each has a calculated average fragment size of within about 150 bases, more often about 100 bases, of the calculated average fragment size of the other). Because PCR is generally more efficient for shorter fragments, the use of fragments of similar length also ensures non-biased PCR amplification between fragments resulting from digestion with different enzymes at subsequent steps in library construction. The calculated average fragment size produced by digestion of a particular sample with a particular enzyme can be determined in a variety of ways. In one embodiment, a database of mRNA/cDNA sequences corresponding to a selected class of mRNAs is used as a representative proxy for the entire population of mRNAs of that class. One database suitable for this analysis is Genbank (accessible at, e.g., http://www.ncbi.nlm.nih.gov/). Using this method, a set of

mRNA sequences known to be expressed in a specified tissue (e.g., brain), organism (e.g., rat, human), or phylum (e.g., mammalia) are identified. Such identification may be easily accomplished because sequences in databases such as Genbank are annotated, so that an investigator can select sequences with particular properties. The frequency and distribution of particular restriction enzyme recognition sites in the selected population of sequences is then determined, e.g., by inspection, but most conveniently by using a computer program such as GCG (Genetics Computer Group Inc., Madison, WI) or Sequencher (Gene Codes Corp, Ann Arbor, MI). In addition, the distribution of restriction sites in the population can be determined using publicly available computer software, and enzymes that frequently cut at clustered sites identified; such enzymes are less desirable than those that recognize more evenly distributed sites.

Table II summarizes an experiment in which enzymes suitable for use with dscDNA prepared from rat mRNA were identified. To identify these enzymes, a collection of 489 full-length rat mRNA/cDNA sequences was collected from Genbank. The selected were from rat, included a polyA-signal at 3' end as well as an entire protein coding sequence (ORF) and at least 100 basepairs of 5' UTR. The mRNAs sequences analyzed had an average mRNA length of 2257 bases (and an average coding sequence length 1509 bases and average 3' untranslated region of 604 bases). The restriction pattern predicted for digestion of this polynucleotide set was determined using the GCG program described *supra.*

Exemplary enzymes for digestion of mammalian sequences include Alu I, Cvi RI, Dpn I, Hae III, Rsa I, Cvi J1 and Tha 1 are examples of enzyme that are not preferred, based on the average calculated fragment size for these enzymes. As is apparent from the table, most suitable enzymes recognize 4-base restriction sites and are blunt-cutters. As determined in the experiment summarized in Table II, preferred combinations of enzymes for construction of libraries from mammalian sequences are Dpn I and Rsa I, because they produce fragments of similar size in the desired size range.

TABLE II

Enzyme	Recognition site	Rec. sites/ mRNA	Not cleaved	Average size
Alu I	AGCT	13.07	0	175
Cvi JI	RGCY	51.89	0	47
Cvi RI	TGCA	11.36	3	199
Dpn I	GATC	07.17	13	**319**

Hae III	GGCC	13.23	0	216
Rsa I	GTAC	05.21	24	**424**
Tha I	CGCG	02.70	171	1044

In alternative embodiments, the average fragment size can be determined empirically. For example, average fragment size can be determined by PCR amplification of large number (e.g., at least 500) of clones from a normalized or subtracted library with vector-specific primers, followed by size determination of inserts on agarose gels.

As noted above, each restriction digestion is carried out separately (i.e., in a separate reaction vessel). Table III provides a flowchart illustrating the production of restriction digested dscDNA from a tissue pair using restriction enzymes Dpn 1 and Rsa 1. Parenthetical numbers are used to refer to specific products (i.e., reagents) produced or used for library production.

TABLE III

(normal) tissue → a) Dpn 1 digest (1)
b) Rsa 1 digest (2)

(diseased) tissue → a) Dpn 1 digest (3)
b) Rsa 1 digest (4)

In embodiments in which digestion is carried out with a single enzyme, any enzyme that would have been suitable as part of an enzyme pair may be used (e.g., Dpn 1 or Rsa 1).

C. Addition of Adaptors

According to the invention, the digested fragments (e.g., digests 1-4 in Table III) are divided into two aliquots and each aliquot is ligated to an adaptor oligonucleotide, i.e., the first aliquot is ligated to a first adaptor and the second aliquot is ligated to a second adaptor. The adaptors used are usually designed to create a 22 to 40 base upper strand hybridized to a 8-12 base lower strand (i.e., partially double-stranded). Adaptors are ligated to dscDNA fragments using methods well known in the art. For example, unphosphorylated oligonucleotides may be ligated to dscDNA fragments in a

standard ligation reaction (e.g., a buffered mixture containing adaptors, fragments, 0.3 mM ATP and T4 DNA ligase, incubated for 12h at 14°C).

The adaptors are designed according to the following criteria:

1) The ligation of the adaptor to the fragment should reconstitute the restriction enzyme recognition sequence for the restriction enzyme used to produce the fragments;

2) The adaptor should have a sequence sufficiently long and complex to serve as targets for amplification by the polymerase chain reaction (PCR), e.g., nested PCR.

3) The first and second adaptors should have different sequences so that a molecule containing both adaptor sequences at opposite ends of a fragment can be differentiated from a molecule containing the same adaptor sequence at each end by PCR amplification using suitable primers.

Methods for preparation of normalized and subtracted libraries by using adaptors suited to PCR amplification are known in the art and may be referred to in the practice of the present invention. See, e.g., Straus and Ausubel, 1990, *Proc. Nat'l Acad. Sci.* 87: 1889; and Diatchenko et al., 1996, *Proc. Nat'l Acad. Sci.* 93:6025-30; see also U.S. Pat. No. 5,759,822, all of which are incorporated herein by reference.

Exemplary adaptors are shown in Table IV, along with primer sets that may be used for PCR amplification:

Table IV

No	first adaptor	second adaptor	Corresponding primers
1*	5'-CTAATACGACTCACTATAGGGCTCGAGCGGCCGCCCGGGCAGGT-3' 5'- ACCTGCCCGG-3'	5'-CTAATACGACTCACTATAGGGCAGCGTGGTCGCGGCCGAGGT-3' 5'- ACCTCGGCCG-3'	5'-CTAATACGACTCACTATAGGGC-3'; Nested PCR Primer 1: 5'-TCGAGCGGCCGCCCGGGCAGGT-3'; Nested PCR Primer 2: 5'-AGCGTGGTCGCGGCCGAGGT-3'
2*	5'-TCGAGCGGCCGCCCGGGCAGGT-3' 5'- ACCTGCCCGG-3'	5'-AGCGTGGTCGCGGCCGAGGT-3' 5'- ACCTCGGCCG-3'	5'-TCGAGCGGCCGCCCGGGCAGGT-3' 5'-AGCGTGGTCGCGGCCGAGGT-3'

*partially double-stranded.

Table V provides, in schematic terms, a flowchart illustrating the addition of adaptors to the products of Table III. In the illustration, the first adaptor is designated "Adaptor A" or "Adaptor C," and the second adaptor is designated "Adaptor B" or "Adaptor D," with different first and second adaptors being used for fragments produced using different restriction enzymes. Although pairs such as A and C or B and D will have different sequences at the end ligated to the fragment (so that the appropriate restriction fragment is regenerated upon ligation) to the extent possible the adaptors are designed to share the same sequence, e.g., to facilitate subsequent PCR amplification.

Table V

(normal) tissue →	a) Dpn 1 digest (1) →	i) adaptor A (1A) ii) adaptor B (1B)
	b) Rsa 1 digest (2) →	iii) adaptor C (2C) iv) adaptor D (2D)
(diseased) tissue →	a) Dpn 1 digest (3) →	i) adaptor (3A) ii) adaptor B (3B)
	b) Rsa 1 digest (4) →	iii) adaptor C (4C) iv) adaptor D (4D)

The adaptor-ligated fragments corresponding to each of the separate digestion reactions may be, and typically are, combined before proceeding to the subsequent subtraction and normalization protocols. For example, referring to Table V, *supra*, 1A + 2C, 1B + 2D, 3A + 4C, 3B + 4D may be combined if adaptors A and C and adaptors B and D differ only at the 3' end (in order to reconstitute the restriction site). However, if desired, the reactions may be combined at later stages, or, alternatively, they may be kept separate.

D. Production of Subtracted libraries

Subtracted libraries (i.e., normalized-subtracted libraries) are used to efficiently identify genes that are differentially expressed in a pair of tissues. Two subtracted libraries are produced, a "driver-subtracted" library and a "tester-subtracted library." When the "tester tissue" is stimulated tissue and the "driver tissue" is unstimulated, the "driver-subtracted" library will be enriched for genes down-regulated by stimulation and the "tester-subtracted" library will be enriched for genes up-regulated by stimulation.

Methods for normalization, substraction and simultaneous normalization and subtraction are known (see, e.g., Ausubel §§5.8-5.9 and discussion *infra*). In one embodiment, the normalized-subtracted libraries of the invention are made essentially according to Diatchenko et al. *supra*. In another embodiment, the production of the normalized-subtracted libraries includes the following steps:

i) First Annealing Step

The following mixtures of adaptor-free digested fragments and adaptor-linked fragments are prepared and annealing reactions carried out (Table VI). The adaptor-free fragments are added in excess over the adaptor-linked fragments, e.g., at an about 20:1, 10:1, or 5:1 ratio. Multiple ratios can be used.

Table VI

driver-subtracted	tester-subtracted
Rxn 1) anneal 1A + 3	Rxn 5) anneal 3A + 1
Rxn 2) anneal 1B + 3	Rxn 6) anneal 3B + 1
Rxn 3) anneal 2C + 4	Rxn 7) anneal 4C + 2
Rxn 4) anneal 2D + 4	Rxn 8) anneal 4D + 2

The mixture is heat-denatured and allowed to anneal, e.g., by heat-denaturation for 90 seconds at 99°C followed by incubation at 68°C to allow annealing in 1 M NaCl, 50 mM HEPES (pH 8.3) and 4 mM Cetyltrimethylammonium bromide. Annealing is allowed to proceed to multiple different Cot values by incubating samples or aliquots for varying times (e.g., 4-12 h for a first sample and 10-24 h for second sample). Hybridization to multiple Cot values results in a more completely normalized library and/or increases the likelihood of enrichment of all differentially regulated genes. It will be recognized that in

the annealing step, abundant sequences represented in the adaptor-ligated population will become double stranded most rapidly, so that, as to adaptor-ligated single-stranded molecules, the library become enriched for low-copy number molecules present in the adaptor-ligated population. When annealing to multiple Cots is carried out, the products can be combined prior to the second annealing step, *infra*, or, alternatively, can be maintained separately throughout the amplification and optional cloning steps.

ii) Second Annealing Step

The reactions mixtures of Table VI, *supra*, are combined and allowed to undergo a second hybridization step with excess (e.g., an about 20:1, 10:1, or 5:1 excess) freshly denatured driver (i.e., adaptor-free fragments), as shown in Table VII.

Table VII

driver-subtracted
Rxn 9) products of Rxns 1 + 2 + additional denatured fragment 3*
Rxn 10) products of Rxns 3 + 4 + additional denatured fragment 4

tester-subtracted
Rxn 11) products of Rxns 5 + 6 + additional denatured fragment 1
Rxn 12) products of Rxns 7 + 8 + additional denatured fragment 2

*(see Tables III and VI)

Annealing is allowed to proceed to different Cot values by incubating samples or aliquots for various times (e.g. 4-20 h).

iii) Amplification

After hybridization, PCR amplification is performed to isolate sequences of interest. In general, only molecules carrying adaptors at both ends can be amplified exponentially by PCR. Other species carry one adaptor at one end and are amplified with linear kinetics, whereas non-adaptor-ligated molecules are not amplified at all. Thus, the double adaptor-ligated population enriched in low-abundance or differentially expressed genes is isolated by PCR amplification. Typically, PCR amplification is done in a 2-step protocol using nested primers for the second amplification.

E. Production of Normalized Libraries

Normalization is the process by which redundant clones in a library are removed, without reducing the complexity of the library. After successful normalization, approximately equal numbers of all expressed genes are present in a library.

Typically normalization methods are based on reassociation kinetics of re-annealing of nucleic acids in which denatured DNA is hybridized to an excess amount of denatured complementary DNA. Because re-annealing nucleic acids follow approximately second-order kinetics, the most abundant species form double-stranded hybrids most quickly. Thus, at any given Cot, rare or less abundant species will preferentially remain single stranded and abundant species will enter the population of double-stranded molecules. Several methods are available for distinguishing, separating, or differentially amplifying the single stranded species. Exemplary normalization methods are found Soares et al., 1994, Proc Natl Acad. Sci. 91:9228-32; Bonaldo et al., 1996, Genome Res. 6:791-806; and U.S. Patent Nos. 5,637,685; 5846721, 5482845, 5830662, 5702898; and Ausubel, *supra*.

In one embodiment, two normalized libraries (referred to as "tester-normalized and "driver-normalized") are produced. In one embodiment, each normalized library is produced essentially according to the protocol described in §D, *supra*, except that the driver and tester are identical. Thus, in one embodiment, the following reactions in Table VIII are carried out.

Table VIII

driver-normalized	tester- normalized
Rxn 1) anneal 1A + 1	Rxn 5) anneal 3A + 3
Rxn 2) anneal 1B + 1	Rxn 6) anneal 3B + 3
Rxn 3) anneal 2C + 2	Rxn 7) anneal 4C + 4
Rxn 4) anneal 2D + 2	Rxn 8) anneal 4D + 4

It will be appreciated that, if desired, reactions 1 and 2, 3 and 4, 5 and 6, and 7 and 8 can be combined.

III. Optimized Selection of Species for Further Analysis

For each library produced, further analysis is carried out to identify sequences likely to be of particular interest. These include genes in the low abundance classes from normalized libraries and differentially expressed genes.

The combination of screening both normalized as well as normalized-subtracted libraries allows comprehensive analysis of the actual expression status of the material under investigation. Previous methods for gene expression analysis operating on a large set of genes (cDNA arrays, oligonucleotide arrays), which require the *a priori* knowledge of the genes under investigation and are considered to be "closed" systems. In contrast, the method disclosed herein combines high-throughput methods for identification of rare or differentially expressed genes, but also permits analysis with no prior knowledge about the gene expression changes expected. That is, the genes under investigation are generated by the method itself and are usually significantly more relevant for the biological process than a preselected set of genes.

A. Generally

In one embodiment, the preferentially amplified or cloned products of subtraction, normalization or combination subtraction-normalization methods are obtained, as described above or by other methods of normalization and/or subtraction. The resulting cDNA (libraries) are subcloned by ligation into a vector capable of propagation in a bacterial or eukaryotic cell. Typically, the clones are propagated in bacterial cells. A number of suitable vectors and cloning methods are known (see, e.g., Sambrook, and Ausubel, both *supra*), including "TA" cloning of PCR products (Stratagene, La Jolla, CA) or blunt-end ligation into a vector of fragments following a fill-in reaction using T4 DNA polymerase and dNTPs.

Further analysis is then carried out by propagating a large number of clones (i.e., by growing a large number of colonies or plaques containing clones from the library(s)). Typically, at least about 5000 clones, more often 10,000, sometimes 15,000 and frequently 25,000 clones are propagated. Because of the large number of clones that are analyzed, it is most convenient and practical to grow clones in multiwell plates (e.g., 384-well plates), using robotic means for growing and picking colonies. Suitable means are known in the art and are described at, e.g., Nguyen et al., 1995, *Genomics* 29:207-216. Alternatively, large numbers of clones can be grown and picked manually.

The insert (i.e., cloned sequences) from each of the clones is isolated and positioned on an array for further analysis. That is, the insert DNAs are immobilized at identified positions in a matrix suitable for hybridization analysis. In one embodiment, high-density filter arrays (HDFA) containing up to 12,000 PCR products per 8x12 cm

membrane are used (Nguyen et al, *supra*). Alternatively, sequences may "printed" onto glass plates, as is described generally by Schena *et al.*, 1995, *Science* 270:467-470.

Most conveniently, the insert corresponding to each clone is amplified by PCR using vector specific primers for spotting on the array. However, other approaches can be used. For example, DNA from each clone can be isolated, the DNA can be digested with a restriction enzyme(s) that cuts at the boundary of the vector and insert, and the insert sequence can be isolated and spotted on the array.

The arrayed sequences are then probed with labeled cDNA derived from "driver" (e.g., unstimulated) tissue or "tester" (e.g., stimulated) tissue. Labeled probes can be prepared using methods known in the art, e.g., by reverse transcription of isolated RNA from the driver and tester tissues in the presence of radiolabeled or fluorescently-labeled nucleotides (see, e.g., Ausubel, *supra*; Kricka, 1992, Nonisotopic DNA Probe Techniques, Academic Press San Diego, CA.; Zhao *et al.*, 1995, *Gene* 156:207; Pietu *et al.*, 1996 *Genome Res.* 6:492). Alternative methods for preparing probes, e.g., riboprobes, are well known and their use is contemplated in some embodiments of the invention.

Optimal hybridization conditions for probing will depend on the type of array (e.g., filter, slide, etc.) selected, the method of labeling probe, and other factors. Hybridization is carried out under conditions of excess immobilized (arrayed) nucleic acid. General parameters for specific (*i.e.*, stringent) hybridization conditions for nucleic acids are described in Sambrook and Ausubel. Suitable hybridization conditions for probing high density arrays are provided in Shena *et al.*, 1996, *Proc. Natl. Acad. Sci. USA,* 93:10614, and Nguyen, *supra*.

When fluorescently labeled probes are used, the fluorescence emissions at each site of a transcript array are detected (e.g., by scanning confocal laser microscopy or laser illumination, see, e.g., Shalon *et al.*, 1996, *Genome Research* 6:639-645; Schena *et al.*, 1996, *Genome Res.* 6:639-645; Ferguson *et al.*, 1996, *Nature Biotech.* 14:1681-1684). When radiolabeled probes are used, autoradiography or quantitative imaging systems (e.g., FUJIX BAS 1000 (Fugi)) may be used. See Nguyen et al., supra, and references cited therein. When it is desirable to determine the ratio of hybridization of two or more probes to the same set of clones, multiple copies of a specific array can be prepared, separately probed, the hybridization intensity be determined for each clone, and a ratio determined. Alternatively, a single array can be repeatedly probed, with washing steps between hybridizations. When differently labeled (e.g., fluorescently-labeled) probes are

used, multiple (e.g., 2) differently labeled probes may be simultaneously hybridized to the same matrix (e.g., rhodamine-labeled driver cDNA and fluorescein-labeled driver cDNA), and, for any particular hybridization site on the transcript array, a ratio of the emission of the two fluorophores can be calculated from simultaneous hybridization to the same array.

One goal of the hybridization is to identify clones corresponding mRNAs expressed at low abundance in driver and tester tissues, particularly clones corresponding to differentially expressed sequences. In the case of normalized libraries, both driver-normalized and tester-normalized libraries are probed with labeled cDNA from the tissue from which they are derived, as indicated in Table IX. Because the signal intensity for any arrayed clone will correspond to the abundance of the corresponding mRNA in the tissue, clones with low intensity signals (i.e., "low signal clones") will correspond to low abundance transcripts (i.e., mRNAs rare in the transcriptome). A "low intensity signal" or "low signal clone" refers to a clone having a hybridization signal in the lowest (e.g., 1st to 20th percentile) or very lowest (e.g., 1st to 5th percentile) range in a ranking of a large number (e.g., 1000) of clone signals in the array. This mRNA class is believed to be enriched for sequences of pharmaceutical importance.

Table IX

Array	Probe	Selection
driver-normalized library array	labeled cDNA probe from driver tissue (e.g., stimulated tissue)	select low signal clones
tester--normalized library array	labeled cDNA probe from tester tissue (e.g., unstimulated tissue)	select low signal clones

There are several advantages to screening both the tester- and driver-normalized libraries. Disease, drug exposure, and other stimulation leads to changes in the overall composition of the transcriptome as well as to transitions of genes from one abundance class into another. Thus, the identity expressed genes as well as their expression levels will be different for the two tissues. These differences will be reflected in the composition of the two libraries both because normalization is never complete (i.e., the resulting library is

never perfectly normalized) and, second, because low abundance genes from one library are sometimes not found in the other.

In the case of the subtracted libraries (i.e., the driver-subtracted and tester-subtracted libraries), both are probed using labeled probes (e.g., cDNA probes) from both RNA sources (i.e. cDNA from driver tissues and cDNA from tester tissues). The ratio between the signals obtained by tester and driver probes indicates the up-regulation or down-regulation of a given clone in response to a stimulus. Thus, probing both driver-subtracted and tester-subtracted libraries will identify all genes that change in expression, either by up-regulation (tester-subtracted) or down-regulation (driver-subtracted). Typically, genes showing at least a 20% (1.2-fold) change is of interest, with genes showing a 2-fold difference in expression are considered to be of particular interest. Preferably, the genes show at least about a 3-fold, 5-fold or 10-fold difference in expression. Clones exhibiting these differences in expression, as detected by hybridization of different probes, are referred to as "high ratio" clones.

Table X

Array	Probe	Selection
driver-subtracted (e.g., enriched for sequences down-regulated in stimulated tissue)	A. labeled cDNA probe from driver tissue	Select a high ratio of A:B Optionally select clones where either A or B give a low intensity signal
	B. labeled cDNA probe from tester tissue	
tester-subtracted (e.g., enriched for sequences up-regulated in stimulated tissue)	A. labeled cDNA probe from driver tissue	Select a high ratio of B:A Optionally select clones where either A or B give a low intensity signal
	B. labeled cDNA probe from tester tissue	

The hybridization analysis described provides an efficient way for prioritizing clones of likely high pharmaceutical significance for further analysis. Selected clones are usually characterized by DNA sequencing and homology analysis. Genes derived from such normalized libraries are used as a representative, relevant and non-redundant gene collection of a particular tissue and a particular biological question for a variety of downstream applications. These genes can serve as targets for array analysis allowing to quantitate gene expression changes in the same or other biological models and

complement the gene collection identified by normalized-subtracted libraries. The analysis of a number of normalized libraries from a variety of central and peripheral tissues under different conditions of stimulation provides an avenue for the ultimate identification of all genes expressed in the species under investigation. In addition, it will be appreciated that, in some embodiments, the arrayed sequences are screened with other probes; for example, an array of sequences differentially expressed in stroke vs. normal brain can be screened with cDNA probe made from mRNA of Alzheimer's Disease brain tissue.

B) "Knock-Down" Analysis

One advantage of the present method is that, among the genes selected for further analysis on the basis of hybridization, the level of redundancy is low (i.e., the number genes that are repeatedly sequenced is low) and the percentage of novel genes detected (genes not previously reported in GenBank) is high.

In contrast, in some prior art DNA libraries contain clones representing a small number of parent genes comprise a large proportion of all the clones in the library. These highly represented (or highly redundant) genes are particularly common in non-normalized libraries, or in libraries from less complex sources, such as specific sub-regions of tissue or cell lines. Random selection of genes from such a library for analysis (e.g. sequencing) results in significant redundancy of effort and expense.

The "knock-down" methods of the invention may be used to further reduce redundancy both in the libraries described herein *supra*, and in libraries prepared by altogether other means (including non-normalized libraries or libraries prepared from specific sub-regions of tissue or cell lines). The knock-down method is used to identify clones that are redundant in a library (i.e., clones generated from transcripts having the same sequence) so that the effort and expense of characterizing the redundant sequences is avoided.

According to the knock-down method, redundant sequences in the library are identified by "prior sampling." That is, prior to the hybridization analysis described in Section III(A), *supra*, or the equivalent of such hybridization, the DNA sequence is determined for representative number of clones, usually at least 50, often between about 100 to about 400 clones, and sometimes more, for example, about 1000 clones. These analyzed clones are referred to as the "prior sample." It is not necessary to sequence the entire clone; rather only one, or optionally both, termini need be sequenced (e.g.,

typically at least about 50 bases are determined, more often between about 200 and 350 bases). The sequences are analyzed, for example by BLAST searching (Altschul et al., 1990, *J Mol. Biol.* 5:403-10). A redundant sequence will appear more often than average: For example, a BLAST-identified sequence appearing as more than 4% of the sample is considered redundant.

In one embodiment of the invention, a set of previously identified genes are included as "knock-down" (e.g., unlabeled) polynucleotide in the "knock-down" method, to identify and avoid further processing of clones that have already been characterized (e.g., sequenced).

If a particular clone or clones is found to be over-represented when compared to other members of the library, DNA may be isolated from the clone(s) (e.g., by PCR amplification of the fragment or insert) and included as an unlabeled (e.g., blocking), or distinctly labeled polynucleotide, during a hybridization of a labeled probe mixture against an array of clones from the library, as described in Section III(A) *supra*. Typically the unlabeled or distinctly labeled "knock-down" polynucleotide is included at a concentration of about 5 to about 100 ng/ml in the hybridization mixture, often from about 5 to about 40 ng/ml. Other useful concentrations will be apparent to one of ordinary skill following the guidance of this disclosure. The unlabeled or distinctly labeled polynucleotides are referred to herein as "knock-down" polynucleotides. In one embodiment, a small number of redundant genes (e.g., one to ten) appearing in the "prior sample" may be included as "knock-down" polynucleotides. In another embodiment, many or all genes appearing in the "prior sample" may be included as "knock-down" polynucleotides.

The included unlabeled (or distinctly labeled) "knockdown" polynucleotide will hybridize to complementary sequences in the labeled probe mixture, reducing the amount of specific labeled probe species available for hybridization to the array. Comparison of the signal of the probe with and without the addition of knockdown polynucleotide will show that the inclusion of the knock-down clone(s) reduces hybridization signals at particular sites on the matrix. The sites of reduced signal correspond to sequences that are represented in the set of "knock-down" polynucleotides (i.e., redundant sequences by frequency or known sequences by prior sampling). Having identified such clones, a decision may be made to not further analyze (e.g., sequence) the clones, saving time and effort.

Alternatively, when the "knock-down" polynucleotides are detectably labeled (using a label that can be distinguished from the probe label), redundant clones will be identifiable by the presence of the distinct signal at the matrix site. This requires an additional labeling step for the "knock-down" polynucleotides and, in one embodiment, requires an additional duplicate hybridization matrix or a measurement of the distinct signal. This is similar to the effort of measuring the signal of the primary (non-knock-down) labeled probe with and without the inclusion of "knock-down" polynucleotides.

Alternately, redundant clones are identified by hybridization of single clones against an array representing the library, rather than by sequence analysis as discussed *supra*. A redundant clone will appear more than once, and more highly redundant clones will tend to appear more than less redundant clones. Non-redundant clones will appear once. In this embodiment, duplications of the array allow testing of as many individual clones as desired to test their redundancy, and the decision may be made to not further analyze (e.g., sequence) the clones, saving time and effort.

IV. ANALYSIS OF METHODS OF LIBRARY CONSTRUCTION

cDNA libraries are a critical reagent used by biologists in the analysis of gene expression and function. Various methods have been used to produce normalized and/or subtracted cDNA libraries (see, e.g., §II *supra* and Ausubel, *supra*). These methods are complex and entail numerous different parameters (e.g., annealing times, polynucleotide concentrations, primer choices, amplification conditions, and the like) and all of which may affect library quality in sometimes unpredictable ways. However, the art lacks a convenient and economical method for evaluating the quality of normalized and/or subtracted cDNA libraries.

As used herein, the "quality" of a subtracted (or normalized-subtracted) library is assessed by the degree to which differentially expressed genes are enriched in the library relative to non-differentially expressed genes. As used herein, the "quality" of a normalized-library (e.g., a tester-normalized or driver-normalized library) is assessed by the degree to which sequences in the library are present in the same abundance.

The present invention provides methods for conveniently assessing library quality. By comparing the quality of libraries made using starting RNA from the same source but made by using different methods, the superior method can be identified (by virtue of producing a higher quality library).

In one embodiment, the method involves making libraries from the same tester and driver RNA but varying parameters. Detectably labeled probe is made from DNA from each library, using standard methods (e.g., nick translation, Ausubel, *supra*). The resulting probes are hybridized to an array of immobilized polynucleotides under conditions of specific hybridization.

Suitable polynucleotide arrays may be produced by any of a variety of methods, but typically are spotted onto glass slides or nylon membranes (e.g., Schena et al., 1995, *Science* 270:467-470, and Zhao et al., 1995, *Gene* 156:207-213). The array is selected to contain at least some polynucleotide sequences representing genes that are differentially expressed in the tester RNA tissue compared to the driver RNA tissue. This may be accomplished generally in two different ways.

In one method, a reference library (e.g. a tester-subtracted library) is produced from tester and driver RNA (e.g., as described *supra*). Typically, the tester and driver RNA used for preparation of the reference library is made from the same tissue sources as used for the libraries to be assessed, although it will be appreciated that this is not strictly necessary. The resulting library is cloned (e.g., by ligation to a vector and transform of bacteria) and DNA corresponding to individual clones prepared (e.g., by PCR amplification using vector primers). DNA from a plurality of the clones (typically at least 50, more often at least 100, more often at least 1000) is applied to a substrate (e.g., glass slide) for hybridization as described *infra*. The resulting cDNAs are spotted onto substrate (e.g. nylon or glass) and the substrate is treated to affix the cDNAs. The array will include differentially expressed sequences (reflecting the library from which the

clones were prepared).

A second method for selection of genes can rely on publications for selection of genes previously reported to be expressed in the tester RNA at higher levels than the driver RNA. These can be identified by their Genbank identifier number, and many can be ordered from commercial sources, and these can be amplified by gene specific primers with PCR.

The resulting arrays are then prehybridized, and hybridized with probe described *supra*. After hybridization (including appropriate washing), the degree of hybridization of each library to various immobilized polynucleotides is detected and compared (e.g., the detectable signal is quantitated). As shown in the Examples, and in Figures 2-4, the intensity of hybridization of the labeled probe to an immobilized

polynucleotide in the array is indicative of the relative abundance of the probe sequence in the library. For example, the more enriched a library is for a differentially expressed gene, the greater the intensity of the hybridization of probe from that library to the immobilized gene sequence.

According to the invention, a higher quality library is identified because at least one differentially expressed sequence shows higher hybridization signal (compared to a library of lower quality). More often, a higher quality library is characterized by a higher hybridization signal to a plurality of different differentially expressed genes on the array, e.g., at least about 5, 10, 20 or 30 sequences or at least about 5%, 10% or 50% of the genes on the array that are differentially expressed (i.e., show an at least 1.2-fold, preferably an at least 2-fold, often at least 3-fold difference in expression between the tester and driver RNAs). If the differentially expressed sequence is rare (i.e. expressed at a low level relative to the average sequence expression level), the hybridization signal of the rare sequence in the improved subtracted-normalized library will increase relative to a tester-subtracted library. Conversely, if a differentially expressed sequence is abundant (i.e. expressed at a higher level relative to the average sequence expression level), the hybridization signal of the abundant sequence in the improved subtracted-normalized library will decrease relative to a tester-subtracted library. Thus, the method provides for the detection of rare clones that are differentially expressed between two conditions.

V. EXAMPLES

EXAMPLE 1
Use of "Knock-Down" Method

A microglia cell line was stimulated with lipopolysaccharide (LPS, 100 ng/ml) and interferon-γ (IFN-γ,100 U/ml) in a culture dish. Stimulated and unstimulated cells were harvested at 12 hours and a tester-subtracted library prepared (SL18). In this specific case, Rsa I was used as restriction enzyme to digest The tester and driver dscDNAs were digested with Rsa I, and adaptor set 1 (see Table IV, *supra*) was used for tester ligations. The first and second hybridizations were for 8 and 16 h, respectively. PCR amplification (primary PCR: 25 cycles, secondary PCR: 12 cycles) was with primer set 1, and products were cloned in pCR 2.1. Primer set 1 is shown *supra* in Table IV.

To identify sequences useful in the knock-down protocol, randomly chosen clones were submitted for DNA sequencing and sequence results were analyzed using the BlastN algorithm. Of 134 sequences identified by BlastN there were a number of genes represented more than once. Four unique genes were represented multiple times by 5, 5, 5, and 6 redundant clones, respectively, accounting for more than 15% of the BlastN identified sequences. "Knock-down" hybridization matrix analysis proceeded with using these genes as "knock-down" polynucleotides. Another 6,000 colonies from the library was picked, and amplified inserts were arrayed on nylon membranes in triplicate. Membranes were each hybridized to ^{32}P-labeled tester and driver cDNAs under stringent conditions, signal intensities analyzed by phosphoimaging and ratios of signal intensities calculated.

"Knock-down" of labeled tester cDNAs hybridization signal intensity was accomplished by inclusion of unlabeled "knock-down" polynucleotides during probe denaturation prior to hybridization. As shown in Figure 1, inclusion of the knock-down polynucleotides resulted in a reduction in signal for redundant clones. In this library, "knock-down" analysis identified 610 clones as redundant, and further analysis (e.g., sequencing) of these genes was thus avoided.

Clones showing at least a 2-fold difference in signal intensities between tester and driver were selected for DNA sequencing and further analysis. Out of the 6,000 original clones in the library, for SL 18 a total of 384 differentially regulated clones were identified. The results of sequence analysis of these clones up-regulated by LPS/IFN-γ, is shown in Table XI:

Table XI*

Library	Known Genes	Similar Genes	Unknown Genes
SL18	52 %	22%	26%

*Gene classification is based on BlastN results using the most recent version of Genbank as database. Genes are considered to be "known" if they display a high degree of similarity (>80% identity on nucleotide level)) to a database entry, as similar if they display a distant similarity (40-80% identity on nucleotide level) and as unknown if they do not show any homology or an insignificant homology to a database entry.

The identification, in this experiment, of redundant clones demonstrates the utility of this method for efficient high-throughput analysis of a large number of genes. In addition, the

large number of unknown genes identified is a further validation for the completeness of the analysis.

EXAMPLE 2

Knockdown Selection of Redundant Clones

A mouse microglial cell line known to respond to stimulation by incubation in media containing lipopolysaccharide (LPS) and gamma interferon (γIFN) was used. mRNA was purified from cells before (= driver) and after stimulation (=tester). A normalized and subtracted cDNA library was prepared and cloned in bacteria ("Library 1").

For a representative number of clones (670), sufficient sequence was determined to assign a Genbank identifier tag (GID) based on a BLAST comparison. Clones matching a GID for MERANTES (GID X70675) were highly represented in the sample (10 clones of 670, or approximately 1.5%). DNA corresponding to the MERANTES sequence was amplified by PCR to produce "knockdown cDNA."

Radiolabeled cDNA probes were prepared from approximately 0.5 micrograms of tester or driver mRNA. The knockdown cDNA was boiled 5 minutes, cooled on ice, and approximately 1 microgram was added to aliquots of Radiolabeled tester probe. Equivalent aliquots of Radiolabeled tester probe and driver probe were used without the addition of knockdown cDNA. The probe or probe/knockdown mixtures were incubated at 68°C for 20 minutes and hybridization solution 50% formamide, 5 X SSC, 5X Denhardt's reagent, 1 % SDS, 0.025% sodium pyrophosphate) was added.

Each of the probe mixtures was hybridized to nylon membranes onto which PCR-amplified cDNA prepared from the 670 partially sequenced clones from Library 1 had been spotted. Hybridization was for 20 hours at 42°C and was followed by washing and signal detection.

Quantitation of the signal level of tester, knockdown-tester and driver hybridizations allowed the selection of clones upregulated by LPS and γIFN, based on their tester/driver ratios. Further, the signal ratio of tester/knockdown-tester allowed for the identification of clones that match the knockdown cDNA. All 10 clones corresponding to MERANTES were identified by an elevated tester/knockdown-tester ratio, with an average tester/knockdown-tester signal ratio of 6.4 fold (stdev 2.2). In contrast, the average tester/knockdown-tester signal ratio for all clones was 1.38 (stdev

0.7). There was one clone with tester/knockdown-tester ratio above 3 fold that was not MERANTES. The selection and effort of further handling of redundant clones (e.g. MERANTES) can be reduced by rejection of clones having an elevated tester/knockdown-tester ratio (e.g. greater than 3)

EXAMPLE 3

Improved Method for Evaluating Quality of Normalized and Subtracted cDNA Libraries

A. Preparation of Tester and Driver ds cDNA

Human fibroblasts (ATCC CRL 2091) were grown to approximately 60% confluence in 15 cm Petri dishes in Dulbecco's Modified Eagle Medium (DMEM), 10% Fetal Calf Serum (FCS). The cells were washed 3 times with DMEM lacking FCS. After a 48 hour incubation in DMEM with 0.1% FCS the medium was replaced with fresh medium containing 10% FCS (serum stimulation). Cells were collected at two different time points. One batch of cells was collected just prior to serum stimulation (serum stimulated cells). This sample served as a time zero reference from which "driver" RNA was prepared. Another batch was collected 6 hours after the addition of FCS. This sample served as a stimulated sample from which "tester" RNA was prepared (serum starved cells).

Total RNA from these samples was prepared using Trizol (Life Technologies). mRNA was selected using Oligotex Kit (Quiagen). The poly A^+ RNA was reverse transcribed using an Oligo dT priming method and converted into double strand cDNA (ds cDNA) using standard methods.

B. Preparation of Normalized and Subtracted Libraries

The ds cDNA was digested with Rsa I (NEB). The Rsa I-digested tester and driver ds cDNA were divided into two aliquots each, and each aliquot was ligated to an adapter oligonucleotide (Adapter set No. 1, shown in Table IV, *supra*). The ligation reaction was performed for 12 hours at 16°C using T4 DNA Ligase (2000 U/μl).

Normalized-subtracted and normalized libraries were prepared as described in § D and E, *supra* , respectively,, using different tester/driver ratios and different conditions for the two annealing steps, as summarized in the table below

Library ID	Library Description	Ratio Tester/ driver	Annealing time (First annealing step)	Annealing time (Second annealing step)
A	Driver-Normalized	1:5	9 hours	18 hours
B	Tester-Normalized	1:5	9 hours	18 hours
C	Normalized-Subtracted , Tester-Subtracted	1:5	9 hours	18 hours
D	NORMALIZED-SUBTRACTED, Tester-Subtracted	1:15	9 hours	18 hours
E	NORMALIZED-SUBTRACTED, Tester-Subtracted	1:10	9 hours	18 hours
F	NORMALIZED-SUBTRACTED, Tester-Subtracted	1:10	12 hours	18 hours
G	NORMALIZED-SUBTRACTED, Tester-Subtracted	1:10	12 hours	36 hours
H	NORMALIZED-SUBTRACTED, Driver-Subtracted	1:20	9 hours	18 hours
I	NORMALIZED-SUBTRACTED, Driver-Subtracted	1:10	9 hours	18 hours
J	NORMALIZED-SUBTRACTED, Driver-Subtracted	1:10	12 hours	18 hours
K	NORMALIZED-SUBTRACTED, Driver-Subtracted	1:10	12 hours	36 hours

Following annealing, a 2-step (nested) PCR amplification was performed to isolate sequences of interest. In the first PCR reaction only molecules which different adapter sequences on each end are amplified exponentially by the adapter-specific primer PCR1. The number of PCR cycles needed to obtain sufficient amounts of amplicon for analysis depends on the experimental paradigm under investigation, and needs to be determined empirically by performing the PCR amplification procedure with different cycle numbers and analyzing amplicon yields (e.g., by agarose gel electrophoresis) In this analysis, different numbers of PCR cycles (21, 23, 25 and 27) were used for the first PCR amplification whereas the second, nested PCR amplification using nPCR1 and nPCR2 as primers proceeded with 12 cycles for all samples.

PCR primer for first amplification:PCR1, CTAATACGACTCACTATAGGGC

PCR primer pair for second, nested amplification:

nPCR1, TCGAGCGGCCGCCCGGGCAGGT

nPCR2, AGCGTGGTCGCGGCCGAGGT

C. Evaluation of Library Quality

i) Array Preparation

Arrays can be prepared using various materials and protocols (for examples, see Schena, Mark et al., "Quantitative monitoring of Gene Expression patterns with a complementary DNA microarray", Science (1995) v270:467-470, and Zhao, Nanding et al., "High-Density cDNA Filter Analysis: A Novel Approach for Large-Scale, Quantitative Analysis of Gene Expression", Gene (1995) v156:207-213). An array can be comprised of a large number of clonal cDNAs on a substrate. The cDNAs can be produced by various methods, including purification of plasmids and PCR amplification. The cDNAs are commonly attached by treatment with heat, ultraviolet light, chemicals or enzymes, or by reaction with a preactivated surface. One typical array starts with the PCR amplification of 11520 bacterial clones containing cDNAs inserted into a plasmid. These clones are commonly from a normalized-subtracted library and therefor contain genes differential in tester and driver mRNA expression levels. Aliquots of the PCR reactions are spotted onto nylon membrane (Scheicher&Scheull) to produce the array. To this array various standard genes are added, the cDNA fragments are denatured by wetting the membrane in a solution of 0.5M sodium hydroxide, 1.5M sodium chloride to allow better availability for hybridization, neutralized and crosslinked by ultraviolet light (Stratalinker, Stratagene). A particular example of a cDNA array suitable for analysis of

library production methods was prepared. Clones corresponding to 80 genes were selected because their mRNA expression levels in fibroblasts varied upon stimulation by serum, based on cDNA microarray data as described in Iyer, Vishwanath et al., 1999 *Science* v283:83-87, incorporated herein by reference in its entirety for all purposes. Recombinant clones were purchased from Research Genetics and verified by DNA sequencing. The cDNA insert of each clone was PCR-amplified using vector-specific primers. PCR products were verified by gel electrophoresis. PCR products were spotted in sextuplicate on nylon membranes.

ii) Probe Preparation

ds cDNA from each of libraries A-K described *supra* (i.e., the products of the second PCR amplification) were gel purified using a QiaEx Gel purification kit. The purified products were labeled with ^{32}P-dCTP (Klenow, Decamer labeling Kit, Ambion) and unincorporated nucleotides were removed by spin column P30 (BioRad).

iii) Evaluation of Library Quality

The probes were hybridized to the cDNA arrays at 42°C in 5xSSC/50% formamide for 20 hours. The hybridized arrays were washed in 0.1x SSC at 60°C and exposed to phosphorimager screens (Packard Instruments) for approximately 64 h. Hybridization signal intensities were determined by a Cyclone scanner and Optiquant software (Packard Instruments), normalized by controls including genomic DNA standards, and comparisons were made between serum-starved fibroblasts (=driver), serum-stimulated fibroblasts (=tester) and different normalized and subtracted libraries. Signal intensity of filter hybridizations was used to determine the abundance of genes and gene fragments in the material used to make the probe (*see* NUCLEIC ACID HYBRIDIZATION, A PRACTICAL APPROACH, pp. 21-22 and 77-111, Hames BD and Higgins SJ eds., IRL Press (1985), and Kafatos et al., 1979, *Nucleic Acids Research Res.*, 7, 1541).

Analysis of the quantified hybridization signal from the arrays allowed grouping of the arrayed genes into several classes based on signal intensities after hybridization. These classes were called low, medium, or high signal levels (herein, corresponding to clones with approximate signal levels of less than 5000 Digital Light Units or DLU=low, 5000-16000 DLU=medium, greater than 16000 DLU, corresponding

to the intensity of the original radioactive probe hybridized to each spot of cloned cDNA on the array). The arrayed genes were also grouped into classes that increase, maintain, or decrease signal intensity (were regulated in the amount of mRNA produced under condition of tester and driver(e.g., serum-stimulation and serum-starvation). In this example, genes were considered up-regulated if the ratio of their tester/driver signals is greater than 2, genes are considered unchanged if the ratio of their tester/driver signals were greater than 0.85 and less than 1.15, and genes were considered down-regulated if the ratio of their driver/tester signals is greater than 1.5. For example, gene could be of low abundance in driver (i.e. low signal of hybridization, herein less than 5000 DLU) and upregulated (i.e. ratio of tester/driver signals is greater than 2).

In Figures 2-4, selected clones within the different abundance classes illustrate the effect of condition group (Library ID) and PCR cycle length (e.g., 21, 23, 25, or 27 cycles on the representation of the clone in the library. For reference, hybridization values for control (=driver) probe are marked RsaI, 0h, and serum stimulated (=tester) probe are marked RsaI, 6h are included in each graph.

This analysis allowed the determination of enrichment factors for each clone represented on the cDNA array and each normalized and subtracted cDNA library. The enrichment factors describe the change in abundance of a particular gene in normalized and subtracted cDNA libraries and are indicators for the success/quality of that library. The quality of a normalized-subtracted library is assessed by the degree to which differentially expressed genes are enriched in the library. During Tester-Subtracted subtraction, upregulated genes (of abundance higher in tester than in driver) are increased in abundance in the resulting library, and down regulated genes are decreased. During reverse subtraction, the reverse is true (e.g. down regulated genes are increased in abundance in the resulting library). The data show that particular conditions (e.g. F25) can increase further the signal and abundance of low, medium and high abundance genes where their initial abundance are higher in tester than in driver.

The quality of a tester-normalized or driver-normalized library is assessed by the degree to which sequences in the library are present in the same abundance, as assessed by a similar intensity of hybridization to the arrayed clones. In a perfectly normalized library, all of the sequences represented are present in the same abundance. Normalization of the abundance of clones gives a more equal chance of discovering what were initially abundant and non-abundant genes, saving time by reducing redundancy of the clone fragments. The data show that particular conditions (e.g. library B) can increase

further the signal and abundance of low, medium and high abundance genes where their initial abundance are higher in tester than in driver.

The quality of a tester-subtracted normalized library is demonstrated by an increase in the occurrence of genes that are more abundant in tester than in driver, a decrease in the occurrence of genes that are more abundant in driver than tester, and the abundance of genes that remain in the library are normalized. This leads to an increase in the abundance of genes having a low abundance that are more prevalent in tester than driver. The normalization will also decrease the redundancy of very abundant genes that are more prevalent in tester than driver. This effect of normalization will ease the discovery of genes more specific to tester that are rare, and increase the efficiency of identifying all genes in the subtracted library. An equivalent assessment of quality can be made for a driver-subtracted normalized library.

***

For the purposes of clarity and understanding, the invention has been described in these examples and the above disclosure in some detail. It will be apparent, however, that certain changes and modifications may be practiced within the scope of the appended claims. All publications and patent applications are hereby incorporated by reference in their entirety for all purposes to the same extent as if each were so individually denoted.

The disclosure of U.S. Patent Application Serial No. No. 09/365587 entitled "SYSTEM AND METHOD FOR IDENTIFYING CRITICAL REGULATED GENES" (Attorney Docket No. 19488-001000US) filed July 30, 1999, is hereby incorporated by reference in its entirety into this application for all purposes.